DNA uptake sequences in *Neisseria gonorrhoeae* as intrinsic transcriptional terminators and markers of horizontal gene transfer.

ABSTRACT

DNA uptake sequences are widespread throughout the *Neisseria gonorrhoeae* genome. These short, conserved sequences facilitate the exchange of endogenous DNA between *Neisseria* species. Often the DNA uptake sequences are present as inverted repeats that are liable to form hairpin structures. It has been suggested previously that DNA uptake sequence inverted repeats present 3' of genes play a role in *rho*-independent termination and attenuation. However, there is conflicting experimental evidence to support this role. The aim of this study was to determine the role of DNA uptake sequences in transcriptional termination. Both bioinformatics predictions, conducted using TransTermHP, and experimental evidence, from RNA-seq data, were used to determine which inverted repeat DNA uptake sequences are transcriptional terminators and in which direction. Here we show that DNA uptake sequences in the inverted repeat configuration occur in *N. gonorrhoeae* both where the DNA uptake sequence precedes the inverted version of the sequence and also, albeit less frequently, in reverse order. Due to their symmetrical configuration, inverted repeat DNA uptake sequences can potentially act as bi-directional terminators, therefore affecting transcription on both DNA strands. This work also provides evidence that gaps in DNA uptake sequence density in the gonococcal genome coincide with areas of DNA that are foreign in origin, such as prophage. This study differentiates for the first time between DNA uptake sequences that form intrinsic transcriptional terminators and those that do not, providing characteristic features within the flanking inverted repeat that can be identified.

DATA SUMMARY

1. Genome *Neisseria gonorrhoeae* strain NCCP11945 is available in GenBank; accession number: CP001050 (url - http://www.ncbi.nlm.nih.gov/nuccore/CP001050) http://www.ncbi.nlm.nih.gov/nuccore/CP001050

2. RNA-seq data for *N. gonorrhoeae* strain NCCP11945 is available from GEO; series numbers: GSE73032 and GSE58650 (url - http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58650 and http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73032).

I/We confirm all supporting data, code and protocols have been provided within the article or through supplementary data files. X

IMPACT STATEMENT

The human pathogens, *Neisseria gonorrhoeae* and *Neisseria meningitidis* are naturally competent for transformation. These species will take up DNA from the environment, engaging in horizontal gene transfer, which has contributed to the rise in antibiotic resistance in *N. gonorrhoea* and the evolution of virulence factors such as the capsule in *N. meningitidis*, as well as capsule serogroup switching in this species. Uptake of DNA carrying the 10-12 bp neisserial DNA uptake sequence is more efficient than without. Because these sequences are scattered in the genome and are associated with the 3' ends of genes, they have also been associated with transcriptional termination. Here we show that DNA uptake sequences terminate transcription only when other sequence features are present, not in all circumstances. In addition, the presence of these uniquely neisserial sequences can be used as an indicator of horizontal gene transfer.

INTRODUCTION

In *Neisseria gonorrhoeae* and *Neisseria meningitidis*, a short, conserved sequence is scattered throughout the genome. This sequence enhances the uptake of external DNA fragments, providing discrimination that favours DNA acquisition from *Neisseria* spp. over foreign DNA (Goodman & Scocca., 1988). These DNA uptake sequences are present in both commensal and pathogenic *Neisseria* spp. (Marri *et al*., 2010), resulting in horizontal gene transfer not only between different strains of gonococci (Sparling., 1966) but also between the various *Neisseria* spp. (Kriz *et al*., 1999,Spratt *et al*., 1992).

There are several variations on the DNA uptake sequence (DUS), the most abundant of which is the 10 bp 5'-GCCGTCTGAA-3' sequence (Goodman & Scocca., 1988), which occurs almost 2000 times in the *N. gonorrhoeae* strain FA1090 genome. Most of these are part of an extended 12 bp DUS (eDUS; 5'-ATGCCGTCTGAA-3') that has been described to have enhanced uptake efficiency (Ambur *et al*., 2007). Compared to the 10 bp DUS, the 12 bp eDUS has been shown to instigate an around three-fold increase in the rate of endogenous DNA uptake in pathogenic *Neisseria* (Ambur *et al*., 2007). Also identified is a variant of the DNA uptake sequence (vDUS; 5'-GTCGTCTGAA-3'), found more commonly in some of the commensal *Neisseria* spp. (Marri *et al*., 2010). These DNA uptake sequences are scattered throughout the genome, however they are less abundant in regions that may have once

been foreign in origin, such as prophages, but are also less frequent in areas responsible for ribosomal proteins (Bentley *et al*., 2007).

Unlike other repetitive sequences in bacterial genomes, such as transposons and gene duplications, the DNA uptake sequences remain highly conserved. Duplications of sequence will, over time diverge from each other, such that elements such as transposons and duplications accumulate sequence differences between them (Smith *et al*., 1999). However, the short DNA uptake sequences are highly conserved between strains of *Neisseria* spp. and across the various neisserial species, with the most common sequence in the non-pathogenic species varying from the pathogenic sequence by only one base out of 10-12 (Marri *et al*., 2010). It may be that the specificity of the ComP surface protein that recognises the DNA uptake sequence (Cehovin *et al*., 2013) exerts selective pressure to maintain the conserved sequence with high affinity for this receptor. Differences in ComP proteins between the pathogens and non-pathogens and corresponding differences in affinity for their variations in the DNA uptake sequences supports pressure to conserve the sequence (Berry *et al*., 2013).

The DNA uptake sequences of 10-12 bp are also found in the *Neisseria* spp. in an inverted repeat orientation (IR-DUS), often flanking a short central region (Ambur *et al*., 2007,Goodman & Scocca., 1988). These inverted repeat configurations of the DNA uptake sequence have been found downstream or internally of neighbouring genes, many of which meet the criteria to function as *rho*-independent terminators or attenuators (Goodman & Scocca., 1988,Marri *et al*., 2010). The inverted repeats are not thought to further enhance endogenous DNA uptake above that achieved as a single copy of the DNA uptake sequence (Ambur *et al*., 2007), yet the frequent occurrence as inverted repeats suggests that these sequences have an additional function.

Using the genome sequence data from *N. gonorrhoeae* strain NCCP11945 (Chung *et al.,* 2008), the implications of the locations of DNA uptake sequences in this strain are explored. The locations of all of the inverted repeat DNA uptake sequences were identified and their fitness as attenuators and *rho*-independent terminators assessed using both bioinformatics predictions and RNA-seq experimental evidence of transcription across the inverted repeat hairpin loop.

METHODS

Identification of DNA uptake sequence locations. DUS locations were determined using he default settings for Fuzznuc pattern finder (Rice *et al*., 2000) within *x*BASE (Chaudhuri *et al*., 2008) on the *N. gonorrhoeae* strain NCCP11945 genome sequence using the sequences of

the 10 bp DUS (5' – GCCGTCTGAA – 3') (Goodman & Scocca., 1988), 12 bp extended DUS (eDUS; 5' – ATGCCGTCTGAA – 3') (Ambur *et al*., 2007), 10 bp variant DUS (vDUS; 5' – GTCGTCTGAA – 3') (Marri *et al*., 2010) , and 12 bp extended variant DUS (veDUS; 5' – ATGTCGTCTGAA – 3'). Identified DUS were mapped here to the circular genome using Artemis (Rutherford *et al*., 2000) and DNAplotter (Carver *et al*., 2009) (Figure 1). Locations of the Correia Repeat Enclosed Elements (CREE) were determined previously (Snyder *et al*., 2009) and were added to the circular map.

Intrinsic transcriptional terminator identification. IR-DUS locations for the DUS, eDUS, vDUS, and veDUS were determined using the default settings for Fuzznuc pattern finder (Rice *et al*., 2000) within *x*BASE (Chaudhuri *et al*., 2008) on the *N. gonorrhoeae* strain NCCP11945 genome sequence, allowing for a central region of one to ten bases followed by the mirrored DNA uptake sequences consensus (e.g. 5' –GCCGTCTGAA x(1,10) TTCAGACGGC-3'). These were then mapped using the Artemis sequence visualization tool (Rutherford *et al*., 2000). The inverted repeat DNA uptake sequences that could potentially be intrinsic transcriptional terminators were then identified using TransTermHP (Kingsford *et al*., 2007) set to a confidence score of 75/100. Intrinsic transcriptional terminator criteria used in TransTermHP prediction software were based on previous results: the terminator must be -50 to 500 bases 3' of the stop codon (Ermolaeva *et al*., 2000,Kingsford *et al*., 2007); the hairpin stem length must be 10 to 18 bases long with a loop size of 3 to 10 bases (Lesnik *et al*., 2001); and the hairpin must be followed immediately by a poly-T region (Ermolaeva *et al*., 2000,Kingsford *et al*., 2007). If the intrinsic transcriptional terminator is to be functional on both strands it must also be preceded by a poly-A region (Ermolaeva *et al*., 2000,Kingsford *et al*., 2007). Using Artemis (Rutherford *et al*., 2000) and Progressive Mauve genome alignment software default settings (Darling *et al*., 2010), *N. gonorrhoeae* stain NCCP11945 was also scanned for all IR-DUS located within a gene or a maximum of 500 bases 3' of the gene's stop codon, confirming the results of the TransTermHP analysis. A schematic of the criteria used as well as an approximation of the IR-DUS terminator structure is shown in Figure 2.

Ion Torrent RNA-seq. *N. gonorrhoeae* strain NCCP11945 was grown in standard conditions on GC agar (Oxoid) with Kellogg's (Kellogg *et al*., 1963) and 5% $Fe(NO_3)_3$ supplements at 37°C or 40°C in an incubator with 5% CO_2 overnight. Growth from the agar plates was removed using sterile plastic 10 µl loops and immediately placed into 500 µl of RNA*Later* (Life Technologies), mixed, pelleted, and the RNA extracted using the RNeasy kit (Qiagen). RNA quality was determined on the 2100 Bioanalyzer (Agilent) and RNA was only used for RNA-seq when it had a RIN of 9 or above. The Ion Personal Genome Machine, Ion Total RNA-Seq Kit with ERCC RNA Spike-In Control Mix, Ion Express Template kit, and Ion Sequencing 200 kit (Life Technologies) were used to sequence 1 µg of rRNA-depleted RNA.

Data analysis. Following data trimming on the Ion Server for quality, determined against internal Ion Controls and the ERCC RNA Spike-In Controls, the BAM files generated for each

RNA-seq experiment were loaded into NextGENe (v2.3.0 to v2.3.2) for mapping against the reference genome sequence of *N. gonorrhoeae* strain NCCP11945 (CP001050) using default settings. Each of the intrinsic transcriptional terminator sites identified using TransTermHP was manually analysed to assess whether transcripts were present at this location, their orientation, and whether they terminated at or near the DUS site. Transcripts in the location of the DNA uptake sequence were seen as those mapping against this region. The orientation of these reads relative to the consensus genome sequence was indicated within the NextGENe viewer. Reads which ended at or near to the DNA uptake sequence were recorded, taking account of their orientation (e.g. transcription toward the DUS, transcription through the DUS), and the number of reads that ended at the same base position on the genomic sequence. This experimental evidence was recorded for each site and for each RNA-seq experiment (Supplementary Table 1). RNA-seq data was deposited into GEO (accession numbers GSE58650 and GSE73032).

RESULTS and DISCUSSION

Uptake sequences and variant uptake sequences. All of the DNA uptake sequences in *N. gonorrhoeae* strain NCCP11945 were identified (Figure 1), including 445 copies of the 10 bp DUS, 1520 copies of the extended 12 bp eDUS, 120 copies of the variant 10 bp vDUS found more frequently in the non-pathogenic *Neisseria* spp., and nine copies of the extended 12 bp version of the variant veDUS (Supplementary Table 2). Therefore, the majority of the DNA uptake sequences found in *N. gonorrhoeae* strain NCCP11945 fit the consensus from the pathogenic species (94%; 1965 of 2094), as would be expected. There are, however, 129 copies of the vDUS/veDUS most commonly in the non-pathogenic *Neisseria* species (Marri *et al*., 2010), providing further evidence suggestive of horizontal exchange of DNA between the species (Spratt *et al*., 1992). These results are similar to those presented previously for other *Neisseria* spp. genomes, where sequences were also identified using fuzznuc (Frye *et al*., 2013). These results are also in concordance with previous results demonstrating horizontal transfer of sequences containing DNA uptake sequences, including between pathogens and non-pathogens, albeit at a lower rate (Berry *et al*., 2013) and the presence of pathogen DUS in the non-pathogens (Marri *et al*., 2010). It has been suggested that the presence of the DNA uptake sequences ensures that the sequences of conserved regions can be repaired if they accumulate mutations (Davidsen *et al*., 2004), as well as playing a role in the dissemination of new traits within the genus (Snyder *et al*., 2007).

Uptake sequences as inverted repeats and intrinsic transcriptional terminators. Some of the DNA uptake sequence copies are adjacent to one another and present in an inverted repeat configuration. A total of 415 of these inverted repeat instances of paired DNA uptake sequences were found (Table 1, Supplementary Table 3). The majority of these (83%, 345 out of 415) are extended 12 bp DNA uptake sequences followed by its compliment (5' –

ATGCCGTCTGAA –N_n– TTCAGACGGCTA – 3'; 313 copies) or the shorter 10 bp DNA uptake sequence and its complement (5' – GCCGTCTGAA –N_n– TTCAGACGGC – 3'; 32 copies) as described previously (Goodman & Scocca, 1988).

Two-thirds of the DNA uptake sequence inverted repeats have both an upstream gene in range (-50 to 500 bp) and are predicted by TransTermHP to function as intrinsic transcriptional terminators (66%; 275 out of 415) (Table 1). These are roughly equally split between those that are predicted to terminate transcription on both strands (119) and those that are predicted to terminate transcription on one strand (156) (Table 2).

The minority of the inverted repeats of DNA uptake sequences are in a reversed orientation (14%, 57 out of 415), with either a 5' – TTCAGACGGCAT –N_n– ATGCCGTCTGAA – 3' or a 5' – TTCAGACGGC –N_n– GCCGTCTGAA – 3' configuration; none were found for the variant DUS in this strain. These reversed orientation inverted repeats are much less likely to be intrinsic transcriptional terminators; only 25% (14 out of 57) meet both criteria of transcriptional terminators (Table 1). The reverse orientation of the inverted repeated have briefly been mentioned previously in an analysis of DNA uptake sequences in the related species *Neisseria meningitidis*, where 360 pairs (89%) were identified in the standard configuration and 45 (11%) DNA uptake sequences inverted repeat pairs were found in the reverse orientation (Smith *et al*., 1999).

Previously, it was experimentally shown that the DNA uptake sequence inverted repeat within the *dcw* cluster of *N. gonorrhoeae* strain FA1090 does not act as a transcriptional terminator 3' of *mraY* (Snyder *et al*., 2003). This gene lies within the division and cell wall synthesis cluster, where transcriptional termination within the cluster could prove problematic for the expression of its essential genes. The prediction results here agree with the previous experimental evidence. This suggests that the predictions for the other inverted repeats of DNA uptake sequences made here are robust. Further, analysis of the inverted repeat DNA uptake sequences in *N. meningitidis* suggested that some do not have the features to be intrinsic transcriptional terminators, either being within genes or lacking the 3' T's (Smith *et al*., 1999). These previous studies support our starting hypothesis that not all of the inverted repeats of uptake sequences in the *Neisseria* spp. are intrinsic transcriptional terminators.

In a previous study of *N. meningitidis* (Ambur *et al*., 2007), potential transcriptional terminators were identified using GeSTer (Unniraman *et al*., 2002). This software is largely similar to TransTermHP used here (Kingsford *et al*., 2007), in that it searches for inverted repeats likely to form hairpins. GeSTer identified approximately one half of the *N. meningitidis* features as potential intrinsic transcriptional terminators (Ambur *et al*., 2007). In *N. gonorrhoeae* strain FA1090, 91% (361 of 396) of the identified features were predicted to terminate transcription. In addition to the differences in prediction algorithms, some parameters differed including the number of bases between each repeat and the distances

from CDSs. In *N. gonorrhoeae* strain NCCP11945, we have found that 66% (275 of 415; Table 1 and Supplementary Table 3) of all of the inverted repeat DNA uptake sequences identified are predicted by TransTermHP to be intrinsic transcriptional terminators on the basis of hairpin formation, presence of a following poly-T region, and association with the 3' of a CDS. Annotation of CDSs within the genome sequence data therefore influences the criteria for identification of transcriptional terminators. Of the 415 inverted repeats identified here, 189 are between CDSs predicted to be convergently transcribed and 116 of these are predicted to be transcriptional terminators on both strands by TransTermHP (61%).

If the majority of inverted repeat DNA uptake sequences in the gonococcus are able to act as *rho*-independent terminators or attenuators, this would mean that they are potentially able to regulate and/or terminate the very CDS they helped introduce into the *N. gonorrhoeae* genome through enhanced uptake of DUS-containing DNA.

RNA-seq data supports TransTermHP predictions for some IR-DUS as transcriptional terminators. As discussed above, not all of the DNA uptake sequence inverted repeats are predicted to be transcriptional terminators by TransTermHP. For example, the copy within the *dcw* cluster does not have the necessary features to be predicted to terminate transcription by TransTermHP and it has been demonstrated experimentally that there is transcription across this inverted repeat (Snyder *et al*., 2003). Using RNA-seq, it is possible to visualise the transcriptome against the genome sequence and assess the presence of transcriptional termination or, indeed, transcription through the location of a DNA uptake sequence inverted repeat. Using RNA extracted from two different growth cultures under the same conditions, the transcriptome of *N. gonorrhoeae* strain NCCP11945 was investigated in these biological replicates at each of the locations containing an inverted repeat of the DNA uptake sequence where one or both sides of the inverted repeat were predicted by TransTermHP to be intrinsic terminators of transcription. The directionality of the transcript was assessed from the RNA-seq data aligned against the reference genome, in each case looking for convergent transcription toward the inverted repeat to assess if these sequences were acting as bidirectional intrinsic terminators of transcription, even when the annotation did not reflect convergent annotated CDSs. Pooling of the RNA from the two biological replicates or merging of the RNA-seq data before analysis would have made it impossible to identify any differences between the replicates, therefore both datasets were fully analysed and there results compared. Because not all CDSs in the genome are transcribed at any one time, it was expected that there may be no transcript data available for some of the regions under investigation.

Across the two RNA samples grown at 37°C, experimental evidence could be found for transcriptional termination supporting the TransTermHP predictions at 75% (207 out of 275) of these inverted repeat DNA uptake sequence sites (Table 2). Experimental evidence

supporting transcriptional termination was recorded (Supplementary Table 1) indicating the end point of the transcript data relative to the genome data for all transcripts converging on the inverted repeat. Such analysis required RNA-seq data that retained the strand information for the original mRNA transcripts, as can be achieved through Ion Torrent RNA-seq. Therefore, two-thirds of the inverted repeat DNA uptake sequences are predicted to be intrinsic transcriptional terminators and these predictions are supported by RNA-seq data for three-quarters of the locations. An example of the transcript data for one such site is represented in Figure 3a.

To assess if the hairpin structure created by the inverted repeat might be influenced by temperature, RNA-seq data was also collected from *N. gonorrhoeae* grown at 40°C. When this RNA-seq data is included in the analysis, it provides further supporting evidence of transcriptional termination for some locations (Supplementary Table 1). Of those inverted repeat DNA uptake sequence sites predicted to terminate transcription on both strands, the 40°C RNA-seq data contributes evidence supporting the 37°C data, as well as adding three new sites with evidence for one strand (Supplementary Table 1). For each of these three additional sites identified only in the 40°C RNA-seq data, there is no data at the location in the 37°C RNA-seq data, which suggests that either these genes are not expressed at 37°C or that their expression is not captured in either of the two 37°C RNA-seq data sets.

Overall, the bioinformatics prediction and the experimental evidence agree with each other quite well and it can clearly be seen that not all copies of the DNA uptake sequence in *N. gonorrhoeae* present as inverted repeats act as transcriptional terminators. Those that meet the specific criteria set out by TransTermHP, including distance from the gene, formation of a hairpin loop, and presence of poly-T regions (Ermolaeva *et al*., 2000,Kingsford *et al*., 2007), are excellent candidates for transcriptional terminators, but it should not be assumed that all DNA uptake sequences downstream of a gene are acting as such, particularly if they are placed within essential operons (Snyder *et al*., 2003). Indeed, only two-thirds of the IR-DUS (275 of 415) possess the necessary sequence features to act as intrinsic transcriptional terminators and these predictions have been supported by the RNA-seq evidence for 75%. This is in concordance with earlier predictions based on *N. meningitidis* sequence data that suggested that about a quarter of the copies lacked the sequence features to be terminators.

Temperature control of termination. In the two sets of RNA-seq data from *N. gonorrhoeae* grown at 37°C, there is only one example of mRNA that reads through an inverted repeat DNA uptake sequence predicted to be an intrinsic transcriptional terminator and into the adjacent sequence. In this case, at the IR-DUS associated with NGK_1855, an ArsR family regulator, there are both transcripts that end with the inverted repeat and those that go through it (Supplementary Table 1, text in red). There is also a transcript through the IR-DUS in RNA-seq data from cultures grown at 40°C. In total, the RNA-seq data set from *N. gonorrhoeae* grown at 40°C contains 12 examples of mRNA transcripts that go through the

inverted repeat region in a direction predicted by TransTermHP to terminate transcription (Supplementary Table 1, text in red). In addition, at another five locations there is both evidence of termination and read through of a predicted terminator (Supplementary Table 1, text in red). These results suggest that the termination of transcription by the inverted repeat of DNA uptake sequences may be a temperature sensitive phenomenon.

There is an inverted repeat of the 12 bp extended DNA uptake sequence between NGK_0100 and NGK_0101 (Figure 3b), which is predicted to be an intrinsic terminator for both strands and has experimental supporting evidence in both 37°C RNA-seq datasets and support for the NGK_0101 transcriptional direction from the 40°C RNA-seq data. However, the transcript from NGK_0100 goes straight through the inverted repeat in the 40°C sample, as assessed by RNA-seq data that aligns to this region. NGK_0100 is *ileS*, isoleucyl-tRNA synthetase and NGK_0101 is a small hypothetical protein, where the reverse transcription through the CDS may seem inconsequential, since the function and indeed candidacy of this 192 bp annotated feature is uncertain. Yet, the preceding gene NGK_0102 is an *opa* (Figure 3), although the annotation of this has not included the correct initiation codon 5' of the CTTCT repeat region. Temperature sensitive transcriptional read-through at the DNA uptake sequence inverted repeat 5' of *ileS* resulting in anti-sense *opa* NGK_0102 mRNA could attenuate expression of this *opa*. There are differences by mfold analysis of this inverted repeat sequence in the optimal energies and predicted structures between 37°C and 40°C (Figure 4).

There is a second extended DNA uptake sequence inverted repeat copy in this region, between NGK_0101 and NGK_0102 (Figure 3b). It is not predicted to be a terminator on the positive strand. As the annotation of NGK_0101 is on the negative strand, this is 5' of NGK_0101, rather than 3' where a transcriptional terminator would need to be located. It is predicted to be a transcriptional terminator on the negative strand for *opa*, NGK_0102, and there is evidence of transcriptional termination in all three RNA-seq datasets, yet there is also evidence of transcription through the inverted repeat in the 40°C RNA-seq data.

It may therefore be that the function of the intrinsic transcriptional terminators formed by the DNA uptake sequence inverted repeats are influenced by more than just the composition of their surrounding sequence and their position relative to the genes. In another example, NGK_1855 has a DNA uptake sequence inverted repeat that is predicted to be a transcriptional terminator only in one direction, to act only for this NGK_1855, encoding a putative ArsR family transcriptional regulator. The RNA-seq data has transcripts ending within 5 bases of the inverted repeat in the two 37°C datasets. However, for one of the 37°C samples and the 40°C sample there is transcription straight through. The next coding sequence is exodeoxyribonuclease III, which has been shown to be critical for the survival of *N. meningitidis* under conditions of oxidative stress (Silhan *et al*., 2012). It may be that this transcriptional terminator is relaxed in certain conditions, acting as an attenuator rather than a terminator.

Regions of the chromosome lacking CREE and DUS. When mapped to the circular chromosome, three regions (A, B, and C) were identified that lacked the *Neisseria*-specific Correia Repeat Enclosed Element (CREE), DNA uptake sequence, or both repeat elements (Figure 1). Region A is divided into two parts that lack DNA uptake sequences entirely. The first, from 542,500 to 562,000 contains putative prophage genes (Spencer-Smith *et al*., 2012), whilst the second, from 570,000 to 592,000 contains the *mafAB* system and associated *mafB* cassettes (Bentley *et al*., 2007,Parkhill *et al*., 2000). Between these two regions there are 14 copies of the DNA uptake sequence associated with the LOS sialyltransferase gene *lst* (Gilbert *et al*., 1996), one of the eleven *opa* alleles, some metabolic genes, and a number of small hypothetical CDSs. The absence of DNA uptake sequences from the prophage and *mafAB* regions supports their foreign origin as previously reported.

Region B is also made up of two parts, from 832,000 to 850,500 and 863,000 to 882,500 with 4 DUS between. This region is the Gonococcal Genetic Island (GGI) described previously (Dillard & Seifert., 2001,Hamilton *et al*., 2005,Snyder *et al*., 2005), which lacks CREE and has few DUS. This region is also believed to have originated outside of the *Neisseria* spp. (Hamilton *et al*., 2005).

There are no CREE in region C from 1,620,500 to 1,699,000. This spans a region containing lysogenic and filamentous prophage which have been previously outlined in *N. gonorrhoeae* strain FA1090 (Kawai *et al*., 2005,Piekarowicz *et al*., 2006,Piekarowicz *et al*., 2007). Two gaps in DNA uptake sequence density within this region coincide with filamentous prophage NgoΦ6 (bases: 1,651,609 – 1,659,560) and NgoΦ9 (bases: 1,674,315 – 1,681,419). It is also noted that there is a complete lack of DUS and CREE in all filamentous prophage, including: NgoΦ7 (bases: 1,220,249 – 1,228,119) and NgoΦ8 (bases: 1,607,649 – 1,614,013).

Large gaps in DNA uptake sequence density may therefore indicate region of non-*Neisseria* origin, particularly if CREE are also absent.

Conclusion

DNA uptake sequence locations and sequences can be important in understanding both gene expression and sequence origin. Two-thirds of the DNA uptake sequences present as inverted repeats in *N. gonorrhoeae* have all of the sequence features to act as transcriptional terminators. Therefore, their presence within operons cannot be assumed to mediate transcriptional termination without added experimentation, as previously shown within the 20 kb *dcw* cluster (Snyder *et al*., 2003). RNA-seq enables the experimental interrogation of transcriptome data across the genome, supporting predictions and identifying transcriptional terminators. In addition, absence of the *Neisseria*-characteristic repetitive sequences can be a powerful indicator of foreign sequence presence, especially when two *Neisseria*-specific sequence features are both lacking from a genomic region. Through assessment of genomic regions that are not associated with *Neisseria*-specific DNA uptake sequences and Correia repeat enclosed elements, several prophage and regions of horizontal gene transfer can be identified.

ACKNOWLEDGEMENTS

Russell Spencer-Smith and Sabrina Roberts and their work that contributed to this manuscript were previously supported by Kingston University during their MSc by Research and PhD studentships, respectively. Lori Snyder's research is supported by Kingston University.

ABBREVIATIONS

DNA uptake sequence (DUS); inverted repeat DNA uptake sequence (IR-DUS); extended 12 bp DNA uptake sequence (eDUS); variant DNA uptake sequence found in commensal non-pathogenic *Neisseria* spp. (vDUS); extended 12 bp variant DNA uptake sequence (veDUS); inverted repeat extended DNA uptake sequence (IR-eDUS); inverted repeat variant DNA uptake sequence (IR-vDUS); Correia Repeat Enclosed Elements (CREE); and Gonococcal Genetic Island (GGI).

REFERENCES

Ambur O. H., Frye S. A., Tønjum T. (2007). New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. *J Bacteriol* **189**, 2077-2085.

Bentley S. D., Vernikos G. S., Snyder L. A., Churcher C., Arrowsmith C., Chillingworth T., Cronin A., Davis P. H., Holroyd N. E.& other authors. (2007). Meningococcal genetic variation mechanisms viewed through comparative analysis of serogroup C strain FAM18. *PLoS Genet* **3**, e23.

Berry J. L., Cehovin A., McDowell M. A., Lea S. M., Pelicic V. (2013). Functional analysis of the interdependence between DNA uptake sequence and its cognate ComP receptor during natural transformation in *Neisseria* species. *PLoS Genet* **9,** e1004014.

Carver T., Thomson N., Bleasby A., Berriman M., Parkhill J. (2009). DNAPlotter: Circular and linear interactive genome visualization. *Bioinformatics* **25**, 119-120.

Cehovin A., Simpson P. J., McDowell M. A., Brown D. R., Noschese R., Pallett M., Brady J., Baldwin G. S., Lea S. M., Matthews S. J., Pelicic V. (2013). Specific DNA recognition mediated by a type IV pilin. *Proc Natl Acad Sci U S A* **19,** 3065-3070.

Chaudhuri R. R., Loman N. J., Snyder L. A. S., Bailey C. M., Stekel D. J., Pallen M. J. (2008). *x*BASE2: A comprehensive resource for comparative bacterial genomics. *Nucleic Acids Res* **36**, D543-D546.

Chung G.T., Yoo J.S., Oh H.B., Lee Y.S., Cha S.H., Kim S.J., Yoo C.K. (2008). Complete genome sequence of *Neisseria gonorrhoeae* NCCP11945. *J Bacteriol* **190**, 6035-6036.

Darling A. E., Mau B., Perna N. T. (2010). progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS One* **5**, e11147.

Davidsen T., Rødland E. A., Lagesen K., Seeberg E., Rognes T., Tønjum T. (2004). Biased distribution of DNA uptake sequences towards genome maintenance genes. *Nucleic Acids Res* **32,** 1050-1058.

Dillard J. P. & Seifert H. S. (2001). A variable genetic island specific for *Neisseria gonorrhoeae* is involved in providing DNA for natural transformation and is found more often in disseminated infection isolates. *Mol Microbiol* **41**, 263-277.

Ermolaeva M. D., Khalak H. G., White O., Smith H. O., Salzberg S. L. (2000). Prediction of transcription terminators in bacterial genomes. *J Mol Biol* **301**, 27-33.

Frye S. A., Nilsen M., Tønjum T., Ambur O. H. (2013). Dialects of the DNA uptake sequence in *Neisseriaceae*. *PLoS Genet* **9,** e1003458.

Gilbert M., Watson D. C., Cunningham A. -., Jennings M. P., Young N. M., Wakarchuk W. W. (1996). Cloning of the lipooligosaccharide a-2,3-sialyltransferase from the bacterial pathogens *Neisseria meningitidis* and *Neisseria gonorrhoeae*. *J Biol Chem* **271**, 28271-28276.

Goodman S. D. & Scocca J. J. (1988). Identification and arrangement of the DNA sequence recognized in specific transformation of *Neisseria gonorrhoeae*. *Proc Natl Acad Sci U S A* **85**, 6982-6986.

Hagman K. E., Pan W., Spratt B. G., Balthazar J. T., Judd R. C., Shafer W.M. (1995). Resistance of *Neisseria gonorrhoeae* to antimicrobial hydrophobic agents is modulated by the *mtrRCDE* efflux system. *Microbiology* **141,** 611-622.

Hamilton H. L., Domínguez N. M., Schwartz K. J., Hackett K. T., Dillard J. P. (2005). *Neisseria gonorrhoeae* secretes chromosomal DNA via a novel type IV secretion system. *Mol Microbiol* **55**, 1704-1721.

Isabella V. M. & Clark V. L. (2011). Deep sequencing-based analysis of the anaerobic stimulon in *Neisseria gonorrhoeae*. *BMC Genomics* **12**.

Kawai M., Uchiyama I., Kobayashi I. (2005). Genome comparison *In silico* in *Neisseria* suggests integration of filamentous bacteriophages by their own transposase. *DNA Research* **12**, 389-401.

Kellogg D. S. Jr., Peacock W. L. Jr., Deacon W.E., Brown L., Pirkle D.I. (1963). *Neisseria gonorrhoeae*. I. Virulence genetically linked to clonal variation. *J Bacteriol* **85**, 1274-1279.

Kingsford C. L., Ayanbule K., Salzberg S. L. (2007). Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* **8**.

Kriz P., Giorgini D., Musilek M., Larribe M., Taha M. -. (1999). Microevolution through DNA exchange among strains of *Neisseria meningitidis* isolated during an outbreak in the czech republic. *Res Microbiol* **150**, 273-280.

Lesnik E. A., Sampath R., Levene H. B., Henderson T. J., McNeil J. A., Ecker D. J. (2001). Prediction of rho-independent transcriptional terminators in *Escherichia coli*. *Nucleic Acids Res* **29**, 3583-3594.

Lucas C. E., Balthazar J. T., Hagman K. E., Shafer W. M. (1997). The MtrR repressor binds the DNA sequence between the *mtrR* and *mtrC* genes of *Neisseria gonorrhoeae*. *J Bacteriol* **179,** 4123-4128.

Marri P. R., Paniscus M., Weyand N. J., Rendón M. A., Calton C. M., Hernández D. R., Higashi D. L., Sodergren E., Weinstock G. M., Rounsley S. D., So M. (2010). Genome sequencing reveals widespread virulence gene exchange among human *Neisseria* species. *PLoS ONE* **5**.

McClure R., Balasubramanian D., Sun Y., Bobrovskyy M., Sumby P., Genco C. A., Vanderpool C. K., Tjaden B. (2013). Computational analysis of bacterial RNA-seq data. *Nucleic Acids Res* **41**.

Parkhill J., Achtman M., James K. D., Bentley S. D., Churcher C., Klee S. R., Morelli G., Basham D., Brown D.& other authors. (2000). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**, 502-506.

Piekarowicz A., Majchrzak M., Klyz A., Adamczyk-Poplawska M. (2006). Analysis of the filamentous bacteriophage genomes integrated into *Neisseria gonorrhoeae* FA1090 chromosome. *Polish Journal of Microbiology* **55**, 251-260.

Piekarowicz A., Kłyz A., Majchrzak M., Adamczyk-Popławska M., Maugel T. K., Stein D. C. (2007). Characterization of the dsDNA prophage sequences in the genome of *Neisseria gonorrhoeae* and visualization of productive bacteriophage. *BMC Microbiology* **7**.

Rice P., Longden L., Bleasby A. (2000). EMBOSS: The european molecular biology open software suite. *Trends in Genetics* **16**, 276-277.

Rouquette-Loughlin C. E., Balthazar J. T., Hill S. A., Shafer W. M. (2004). Modulation of the *mtrCDE*-encoded efflux pump gene complex of *Neisseria meningitidis* due to a Correia element insertion sequence. *Mol Microbiol* **54,** 731-741.

Rouquette C., Harmon J. B., Shafer W. M. (1999). Induction of the *mtrCDE*-encoded efflux pump system of *Neisseria gonorrhoeae* requires MtrA, an AraC-like protein. *Mol Microbiol* **33,** 651-658.

Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M. -., Barrell B. (2000). Artemis: Sequence visualization and annotation. *Bioinformatics* **16**, 944-945.

Silhan J. Nagorska K., Zhao Q., Jensen K., Freemont P.S., Tang C.M., Baldwin G.S. (2012). Specialization of an Exonuclease III family enzyme in the repair of 3' DNA lesions during base excision repair in the human pathogen *Neisseria meningitidis*. *Nucleic Acids Res* **40**, 2065-2075.

Smith H. O., Gwinn M. L., Salzberg S. L. (1999). DNA uptake signal sequences in naturally transformable bacteria. *Res Microbiol* **150,** 603-616.

Snyder L. A. S., Cole J. A., Pallen M. J. (2009). Comparative analysis of two *Neisseria gonorrhoeae* genome sequences reveals evidence of mobilization of correia repeat enclosed elements and their role in regulation. *BMC Genomics* **10**, 70-2164-10-70.

Snyder L. A. S., McGowan S., Rogers M., Duro E., O'Farrell E., Saunders N. J. (2007). The repertoire of Minimal Mobile Elements in the *Neisseria* species and evidence that these are involved in horizontal gene transfer in other bacteria. *Mol Biol Evol* **24,** 2802-2815.

Snyder L. A. S., Jarvis S. A., Saunders N. J. (2005). Complete and variant forms of the 'gonococcal genetic island' in *Neisseria meningitidis*. *Microbiology* **151**, 4005-4013.

Snyder L. A. S., Davies J. K., Saunders N. J. (2004). Microarray genomotyping of key experimental strains of *Neisseria gonorrhoeae* reveals gene complement diversity and five new neisserial genes associated with minimal mobile elements. *BMC Genomics* **5**.

Snyder L. A. S., Shafer W. M., Saunders N. J. (2003). Divergence and transcriptional analysis of the division cell wall (*dcw*) gene cluster in *Neisseria* spp. *Mol Microbiol* **47**, 431-441.

Spencer-Smith R., Varkey E. M., Fielder M. D., Snyder L. A. S. (2012). Sequence features contributing to chromosomal rearrangements in *Neisseria gonorrhoeae*. *PLoS ONE* **7,** e46023.

Sparling P. F. (1966). Genetic transformation of *Neisseria gonorrhoeae* to streptomycin resistance. *J Bacteriol* **92**, 1364-1371.

Spratt B. G., Bowler L. D., Zhang Q. -., Zhou J., Smith J. M. (1992). Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. *J Mol Evol* **34**, 115-125.

Unniraman S., Prakash R., Nagaraja V. (2002). Conserved economics of transcription termination in eubacteria. *Nucleic Acids Res* **30,** 675-684.

Vélez Acevedo R. N., Ronpirin C., Kandler J. L., Shafer W. M., Cornelissen C. N. (2014). Identification of regulatory elements that control expression of the *tbpBA* operon in *Neisseria gonorrhoeae*. *J Bacteriol* **196**, 2762-2774.

DATA BIBLIOGRAPHY

1. Roberts, S. & Elbeyioglu F. Gene Expression Omnibus GSE58650 (2014).
2. Snyder, L.A. & Roberts, S. Gene Expression Omnibus GSE73032 (2016).

FIGURES AND TABLES

Figure 1. Circular chromosome map representing the locations of DNA uptake sequences (DUS) and Correia Repeat Enclosed Elements (CREE) in the genome sequence of *N. gonorrhoeae* strain NCCP11945. From the outer circle to the inner circle, each represents the position of features on the positive (outer) and negative (inner) strand: annotated CDSs (red); 10 bp DUS (blue) overlaid with extended 12 bp eDUS (grey); non-pathogen variant 10 bp vDUS (green) overlaid with extended 12 bp versions of the non-pathogen variant sequence veDUS (grey); and CREE (purple). Gaps in these features are labelled A, B, and C. Two gaps in DNA uptake sequence density are present at A, two gaps in DUS density that coincide with a large gap in CREE density are present at B, and C is an area of low CREE density.

Figure 2. Schematic representation of the ideal requirements needed for a potentially functional inverted repeat DNA uptake sequence bi-directional *rho*-independent terminator, using the enhanced 12 bp DNA uptake sequence in the most common orientation as an example.

Figure 3. Transcriptional termination and inverted repeats of the DNA uptake sequence. a) The *mtrRCDE* region contains overlapping promoter elements (bent arrows) that drive transcription of the *mtrCDE* operon (blue CDSs) encoding the MtrCDE efflux pump proteins and the *mtrR* gene encoding the MtrR negative regulator (Hagman *et al*., 1995). MtrR binds to the regulator binding site indicated with the blue/green box (Lucas *et al*., 1997). A Correia Repeat Enclosed Element (beige triangle) sometimes removes this operon from its ancestral regulatory network through insertion between the *mtrR* gene and promoter (Rouquette-Loughlin *et al*., 2004). At the 3’ end of the *mtrCDE* operon is an IR-DUS (red diamond) predicted by TransTermHP and supported by RNA-seq (black bar) to be an intrinsic transcriptional terminator (Supplementary Table 1). The staggered right end of the black bar

represents the ends of transcript from *mtrCDE* toward the IR-DUS from the RNA-seq data. An IR-DUS is also 3' of *mtrA*, which encodes the positive regulator MtrA that also binds to the *mtrCDE* regulator binding site (blue / green box; Rouquette *et al*., 1999). b) Schematic of the chromosomal region containing NGK_0100 *ileS* (orange), NGK_0101 a hypothetical protein CDS (green), and NGK_0102 *opa* (yellow). The IR-eDUS between NGK_0100 and NGK_0101 (red) is predicted to be an intrinsic transcriptional terminator on both strands. The IR-eDUS between NGK_0101 and NGK_0102 (purple) is predicted to be an intrinsic transcriptional terminator on the negative strand but not the positive strand. The staggered ends of the bars below the operon represent the ends of the transcripts from the RNA-seq data, colour coded to match the genes. The red bar and purple bar represent the RNA-seq data from growth at 40°C, where there is transcription through the IR-DUSs.

Figure 4. Boxplots generated by mfold of the region between NGK_0100 and NGK_0101 containing an extended DNA uptake sequence inverted repeat at 37° (a) and at 40°C (b) showing that these differ at the two temperatures. This has an impact upon the predicted structures from mfold, with the three structures predicted at 37°C shown in panels c, d, and e and those at 40° in panels f, g, and h, each in the order of likelihood of the formation of the structure.

Table 1. DNA uptake sequences as inverted repeats in *N. gonorrhoeae* strain NCCP11945.

	Total in genome	Meet both termination criteria[a]
IR-eDUS[b]	313	226
Reverse IR-eDUS[c]	40	11
IR-DUS[d]	32	27
Reverse IR-DUS[e]	17	3
IR-vDUS[f]	13	8
Reverse IR-vDUS[g]	0	n/a
Total	**415**	**275**

[a] On either the positive strand of the genome, the negative strand of the genome, or both strands, the inverted repeat is both within range (-50 to 500 bp) of an annotated CDS from the *N. gonorrhoeae* strain NCCP11945 genome sequence (CP001050) and the sequence is predicted to be an intrinsic transcriptional terminator by TransTermHP (Kingsford *et al*., 2007).

[b] Inverted repeat of the 12 bp extended DNA uptake sequence (5' – ATGCCGTCTGAA –N_{1-10}– TTCAGACGGCTA – 3').

[c] Inverted repeat of the 12 bp extended DNA uptake sequence in reverse orientation (5' – TTCAGACGGCTA –N_{1-10}– ATGCCGTCTGAA – 3').

[d] Inverted repeat of the shorter 10 bp DNA uptake sequence (5' – GCCGTCTGAA –N_{1-10}– TTCAGACGGC – 3').

[e] Inverted repeat of the shorter 10 bp DNA uptake sequence in reverse orientation (5' – TTCAGACGGC –N_{1-10}– GCCGTCTGAA – 3').

[f] Inverted repeat of the DNA uptake sequence variant commonly found in non-pathogenic *Neisseria* spp. (5' – GTCGTCTGAA –N_{1-10}– TTCAGACGAC – 3').

[g] Inverted repeat of the DNA uptake sequence variant commonly found in non-pathogenic *Neisseria* spp. in reverse orientation (5' – TTCAGACGAC –N_{1-10}– GTCGTCTGAA – 3').

Table 2. Predicted intrinsic transcriptional terminators that are supported by RNA-seq.

	Predicted[a]	Supported by RNA-seq[b]	Percentage[c]
Bidirectional[d]	119	37 (bidirectional) & 62	83%
One strand[e]	156	108	69%
Total	**275**	**207**	**75%**

[a] The inverted repeat of the DUS uptake sequence is predicted by TransTermHP to be an intrinsic transcriptional terminator.

[b] The prediction of termination is supported by either or both sets of RNA-seq data from cultures grown at 37°C.

[c] The percentage of predicted transcriptional terminators that have been supported by RNA-seq data to terminate transcription.

[d] The inverted repeat of the DUS uptake sequence is predicted by TransTermHP to be an intrinsic transcriptional terminator on both strands, therefore making a bidirectional intrinsic transcriptional terminator.

[e] The inverted repeat of the DUS uptake sequence is predicted by TransTermHP to be an intrinsic transcriptional terminator on only strand, therefore making an intrinsic transcriptional terminator on either the positive or negative strand of the genomic DNA.